AFRL-IF-WP-TM-2003-1547

PACT: POWER AWARE COMPILATION AND ARCHITECTURAL TECHNIQUES

Prithviraj Banerjee, Alok Choudhary, and Andreas Moshovos

Northwestern University
Center for Parallel and Distributed Computing
2145 Sheridan Road
Evanston, IL 60208

Majid Sarrafzadeh

UCLA
Department of Computer Science
Los Angeles, CA 90095

AUGUST 2003

Final Report for 01 June 2000 – 31 May 2003

Approved for public release; distribution is unlimited.

STINFO FINAL REPORT

INFORMATION DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7334

NOTICE

USING GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA INCLUDED IN THIS DOCUMENT FOR ANY PURPOSE OTHER THAN GOVERNMENT PROCUREMENT DOES NOT IN ANY WAY OBLIGATE THE US GOVERNMENT. THE FACT THAT THE GOVERNMENT FORMULATED OR SUPPLIED THE DRAWINGS, SPECIFICATIONS, OR OTHER DATA DOES NOT LICENSE THE HOLDER OR ANY OTHER PERSON OR CORPORATION; OR CONVEY ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY RELATE TO THEM.

THIS REPORT IS RELEASABLE TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS). AT NTIS, IT WILL BE AVAILABLE TO THE GENERAL PUBLIC, INCLUDING FOREIGN NATIONS.

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

/s/
ANDREW W. HYATT, 2Lt, USAF
Program Monitor

/s/
ROBERT A. EHRET, Chief
Collaborative Simulation Technology & Applications Branch
Information Systems Division
Information Directorate

Do not return copies of this report unless contractual obligations or notice on a specific document require its return.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS**.

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
August 2003	Final	06/01/2000 – 05/31/2003

4. TITLE AND SUBTITLE
PACT: POWER AWARE COMPILATION AND ARCHITECTURAL TECHNIQUES

5a. CONTRACT NUMBER
F33615-00-C-1631

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
69199F

6. AUTHOR(S)
Prithviraj Banerjee, Alok Choudhary, and Andreas Moshovos (Northwestern University)
Majid Sarrafzadeh (UCLA)

5d. PROJECT NUMBER
ARPI

5e. TASK NUMBER
FS

5f. WORK UNIT NUMBER
A3

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Northwestern University
Center for Parallel and Distributed Computing
2145 Sheridan Road
Evanston, IL 60208

UCLA
Department of Computer Science
Los Angeles, CA 90095

8. PERFORMING ORGANIZATION REPORT NUMBER
0650-350-D401

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Information Directorate
Air Force Research Laboratory
Air Force Materiel Command
Wright-Patterson AFB, OH 45433-7334

DARPA/IPTO
3701 Fairfax Drive
Arlington, VA 22203-1714

ONRRO Chicago
536 South Clark Street, Room 286
Chicago, IL 60605-1588

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/IFSD

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-IF-WP-TM-2003-1547

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Report contains color.

14. ABSTRACT

The goal of this project was to take DoD applications written in C and generate power and performance efficient code for systems utilizing the architectural power-aware techniques developed. The PACT project consisted of 3 research tasks: 1) Power-aware architectural approaches, 2) Power-aware compilation strategies, and 3) Power-aware CAD tools for power estimation and synthesis. As part of the power aware architecture research, we developed power aware techniques for on-chip buses, power aware memory hierarchies, and a framework to evaluate heterogeneous embedded systems for performance and energy consumption. As part of the power aware compiler research, we have developed a compiler that takes general C programs and generates power aware codes for three targets: 1) General purpose embedded processor such as the StrongARM, 2) General purpose field-programmable gate arrays (FPGAs), and 3) General purpose application specific integrated circuits (ASICs). We have developed improved strategies for power optimization and management, and improved design methodologies and design philosophies for better estimation and optimization.

15. SUBJECT TERMS
Power Aware Computing and Communication; Power Aware Compilers; Power Aware Architectures

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			2Lt Andrew W. Hyatt
Unclassified	Unclassified	Unclassified	SAR	30	19b. TELEPHONE NUMBER *(Include Area Code)* (937) 904-9162

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

TABLE OF CONTENTS

ABSTRACT

The objective of the PACT project was to develop power-aware compiler, architecture and CAD tool support. The goal of this project was to take DOD applications written in C and generate power and performance efficient code for systems utilizing the architectural power-aware techniques developed. The PACT project consisted of three research tasks: (1) Power-aware architectural approaches. (2) Power-aware compilation strategies (3) Power-aware CAD tools for power estimation and synthesis. As part of the power aware architecture research, we have developed power aware techniques for on-chip buses, power aware memory hierarchies, and a framework to evaluate heterogeneous embedded systems for performance and energy consumption. As part of the power aware compiler research, we have developed a compiler that takes general C programs and generates power aware codes for three targets: (1) General purpose embedded processor such as the StrongARM (2) General purpose field-programmable gate arrays (FPGAs) (3) General purpose application specific integrated circuits (ASICs). Finally, as part of the power-aware CAD tools research, we have developed improved strategies for power optimization and management, and improved design methodologies and design philosophies for better estimation and optimization.

1. TECHNICAL STATUS REPORT

1.1. Goals and objectives for project

The objective of the PACT (Power-aware Architecture and Compilation Techniques) project is to develop power-aware compiler and CAD tool support. The goal is to take DOD applications written in C and generate power- and performance-efficient code for systems utilizing the architectural power-aware techniques developed. The specific goals of the PACT project are to:

(1) Develop novel compiler concepts at various levels that can reduce the total energy consumption in specific applications by factors of 10-100X over conventional, non-power-aware architectures;
(2) Develop compiler techniques to automate the process of generating efficient code that is within a factor of two of the best manual approach with respect to optimizing power under performance and resource constraints;
(3) Demonstrate the usefulness of the compiler concepts on some real applications. We will target specific algorithms/applications that are of interest to DOD

The project URL is at:

http://www.ece.nwu.edu/cpdc/PACT/PACT.html

1.2. Effort of Various PACT Project Tasks and Status

The PACT project originally consisted of three research tasks, architecture, compiler and CAD. However, in January 2002, we were directed by DARPA to terminate our effort in architecture and in the CAD areas with reduced total funding.

Task A: Power-aware architectural approaches
(100% complete)

In this task we developed power aware architectural approaches for:

- Power aware on chip buses
- Power aware memory hierarchies
- A framework to evaluate heterogeneous embedded systems for performance and energy consumption

Task B: Power-aware compilation strategies
(100% complete)

In this task, we have developed a compiler that will take general C programs and generating power aware codes for three targets:

- General purpose embedded processor such as the StrongARM
- General purpose field-programmable gate arrays (FPGAs)

- General purpose application specific integrated circuits (ASICs)

Task C: Power-aware CAD tools for power estimation and synthesis (100% complete)

Research on power aware CAD tools has been pursued in two distinct directions

- Development of improved strategies for power optimization and management
- Development of improved design methodologies and design philosophies for better estimation and optimization

1.3. Quarterly Milestones for Entire Project

We now list the quarterly milestones in a quarter by quarter format for the entire project and how it relates to each task.

The PACT project consisted of two major tasks and extended over 3 years. We provide below a list of milestones that we planned to accomplish in each of the 12 quarters in various tasks.

Sep. 30, 2000:

- Task A: Complete infrastructure/overview for architecture research taking into account feedback from various DARPA application groups (Land Warrior and NASA JPL). Assemble team of students. (COMPLETED)
- Task B: Develop compiler infrastructure for power-aware compiler research (for embedded, FPGAs and ASICs). Assemble team of students. (COMPLETED)
- Task C: Develop CAD tool infrastructure for power-aware CAD research (power estimation and synthesis). Assemble team of students. (COMPLETED)

Dec. 31, 2000:

- Task A: Alternate information encoding schemes for memory controller w/ commodity DRAMs. Memory bus modeling. PACT_sim: Modify Simplescalar simulator for StrongARM instruction set emulation. (COMPLETED)
- Task B: Start modifying SUIF and GCC based compiler framework to generate instructions for StrongARM processor (COMPLETED)
- Task C: Develop logic level power estimator for combinational circuits (COMPLETED)

March 31, 2001:

- Task A: Alternate information encoding schemes for memory controllers w/ modified DRAMs. Memory bus modeling; PACT_sim: Incorporate power models for key processor structures (e.g., caches, TLBs, register file scheduler) into PACT_sim. (COMPLETED)

- Task B: Complete modification of SUIF/GCC compiler to generate StrongARM instructions; integrate with SimplescalarARM simulator. Also develop HDL compiler infrastructure. (COMPLETED)
- Task C: Develop logic level power estimator for sequential circuits (COMPLETED)

June 30, 2001:

- Task A: Power-aware memory refresh/caching optimizations for memory controller.
  PACT_sim: Incorporate power models for key processor structures (e.g., caches, TLBs, register files, scheduler) into PACT_sim. (COMPLETED)
- Task B: Develop PACT HDL compiler to produce RTL VHDL for ASIC and FPGAs (COMPLETED)
- Task C: Develop behavioral level power estimator (COMPLETED)

Sep. 30, 2001

- Task A terminated
- Task B: Generate RTL VHDL from the PACT HDL compiler , integrate with with commercial CAD tools such as Synopsys and Synplicity to generate designs for FPGAs and ASICs (COMPLETED)
- Task C: Complete power estimation tool PACT_estimator at behavioral and logic level. (COMPLETED)
- Demo of PACT toolset version 1.0. Transfer to Accelchip, Synopsys, Synplicity, Cadence and Motorola. (COMPLETED)

Dec. 31, 2001:

- Task B: Develop compiler transformations (temporary variable removal, common subexpression elimination, reverse levelization, pipelining) within the PACT HDL compiler for low power. (COMPLETED)
- Task C: Initial development of low-power synthesis techniques for combinational circuits (PACT-optimize) (COMPLETED)
- Task C terminated

Mar. 31, 2002:

- Task B: Develop compiler algorithms for high-level/low-level compiler transformations (loop interchange, data flow optimizations) within framework of PACT ARM and evaluate using StrongARM. (COMPLETED)

June 30, 2002:

- Task B: Develop algorithms for generating RTL Verilog from PACT HDL. (COMPLETED)

Sep. 30, 2002:

- Task B: Investigate compiler algorithms for technology driven voltage scaling issues, low power Memory Synthesis for ASICs; (COMPLETED)
- Demo of PACT toolset version 2.0. Transfer to Accelchip. (COMPLETED)

Dec. 31, 2002:

- Task B: Develop various compiler optimizations such as Array Access Scalarization on the HDL AST for PACT HDL. (COMPLETED)

Mar. 31, 2003:

- Task B: Develop compiler algorithms for estimating power, area, times for library functions automatically (COMPLETED)

May 31, 2003:

- Task B: Final version of PACT compiler with high- and low-level optimizations. Demo of PACT toolset version 3.0. (COMPLETED)

1.4. Key Accomplishments of this Project (June 1, 2000 to May 31, 2003)

We are going to report on our key accomplishments in three areas:

Task A: Power Aware Architecture
Task B: Power Aware Compiler
Task C: Power Aware CAD

1.5. Task A: Power Aware Architecture

We now report on various research results we have obtained in the area of power-aware architecture.

Power-aware off-chip data bus

Power consumption is becoming increasingly important for both embedded and high-performance systems. Off-chip data buses can be a major power consumer. We present a strategy called “power protocol” that tries to reduce the dynamic power dissipation on off-chip data buses. To accomplish this, our strategy reduces the number of bus lines that need to be activated for data transfer by employing a small cache (called “value cache”)

at each side of the off-chip data bus. These value caches keep track of the data values that have recently been transmitted over the bus. The entries in these caches are constructed in such a way that the contents of both the value caches are the same all the time. When a data value needs to be transmitted over the bus, we first check whether it is in the value cache of the sender. If it is, we transmit only the index of the data (i.e., its value cache address) instead of the actual data value and, the other side (receiver) can determine the data value by using this index and its value cache. Our experimental results using a set of fifteen benchmark codes from embedded systems domain show that power protocol is very effective in practice [3,6], and reduces the bit switching activity on the data bus by as much as 70.7% (with a value cache of 128 entries). We also present results from an implementation that combines our strategy with 1-to-2 encoding, a popular bus encoding strategy for low power. Our results indicate that this combined optimization strategy reduces bit switching activity by 67.8% on the average (across all benchmarks). These reductions in bit switching activity lead to more than 7% reduction on overall system energy on the average for a value cache of 256 entries. We also study the sensitivity of our savings to the value cache capacity and data cache capacity.

Power-aware memory hierarchy

In recent years, both performance and power have become key factors in efficient memory design. We propose a systematic approach to reduce the energy consumption of the entire memory hierarchy. We first evaluate an existing power-aware memory system where memory modules can exist in different power modes, and then propose on-chip memory module buffers, called Energy-Saver Buffers (ESB), which reside in-between the L2 cache and main memory. ESBs reduce the additional overhead incurred due to frequent resynchronization of the memory modules in a low-power state. An additional improvement is attained by using a model that dynamically resizes the active cache based on the varying needs of a program. Our experimental results demonstrate that an integrated approach can reduce the energy-delay product by as much as 50% when compared to a traditional non power-aware memory hierarchy.

A framework to evaluate heterogeneous embedded systems for performance and energy consumption

We present an application-initiated strategy that aims to control the energy consumption, while simultaneously enhancing the performance of a multi-resource heterogeneous embedded system. We assess the benefits of using this strategy by means of a traditional evaluation framework. Even though the overall benefits and improvements are apparent, the performance-energy trade-offs are not prominently noticeable when the traditional framework is used during evaluation. Hence, we propose a framework based on a new metric called energy-resource efficiency (ERE). ERE defines a link between the performance and energy variations in a system. This metric also serves as a guide to determine the amount of resources needed to attain the desired performance and energy behavior. Our experimental results clearly indicate that a heterogeneous system running an application-aware strategy, when correctly calibrated using ERE, leads to great performance and energy gains.

Characterizing and improving energy-delay tradeoffs in heterogeneous communication systems

Communication systems with multiple computing resources are gaining popularity. These devices consume a considerable amount of energy, and require that the system resources be fully utilized at all times. We aim to reduce the energy consumption and also improve the system utilization by incorporating performance-enhancement and power-optimization techniques into a multi-resource heterogeneous communication system. We study the performance and energy improvements contributed by our techniques, using a traditional evaluation framework. However, this framework fails to clearly model the performance-energy tradeoffs in the system. Hence, we use a more relevant framework with a new metric named Energy-resource efficiency (ERE) to study these systems. ERE defines a link between the performance and energy variations in a system to clearly highlight the various performance-energy tradeoffs. Our experimental results show that, when ERE is used to calibrate the performance-energy tradeoffs, one can achieve up to 80% performance and energy gains in a power-aware heterogeneous communication system.

1.6. Task B: Power Aware Compiler

In this research we have developed a compiler that takes general C programs and generating power aware codes for three targets:

- General purpose embedded processor such as the ARM
- General purpose field-programmable gatre arrays (FPGAs)
- General purpose application specific integrated circuits (ASICs)

Development of the PACT Compiler Infrastructure

In this task, we have developed the PACT compiler which solved two problems: (1) it allowed users to develop algorithms in a high level language, namely C, and synthesize hardware designs onto FPGAs and ASICs. (2) it explicitly addressed low power issues during the high-level synthesis stages. The PACT compiler used the SUIF C Compiler from Stanford University as its C front-end. The PACT compiler is a fully modularized three-stage C to HDL Compiler (Figure 1). In the first stage, the C code is parsed into a high-level C type Abstract Syntax Tree (AST). At this stage, both power and performance optimizations may be performed. In general, these optimizations will not require specific information about an HDL representation or about the target architecture, such as precision analysis, constant-propagation, or loop unrolling. During the second stage, the C AST is converted into Finite State Machine (FSM) style HDL AST. Target architecture specific information is inserted into the AST in this phase. Again, power and performance optimizations can be performed. In general, these optimizations require specific information about clocks, cycles, or the target architecture, such as memory pipelining, and clock-gating. Finally, RTL code is generated in the back-end phase. This phase is not designed for optimization; however, if language-specific issues arise they are handled in this phase.

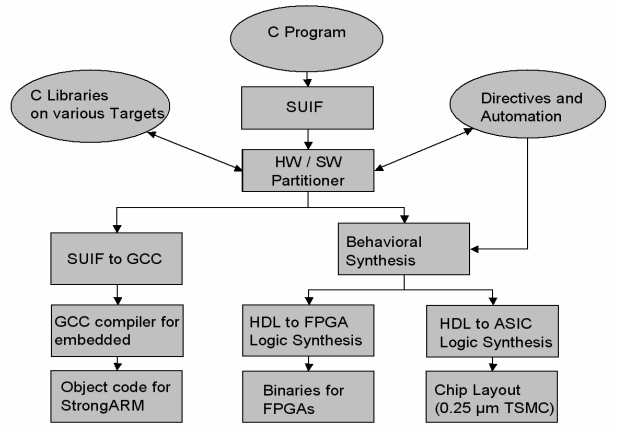

Figure 1: PACT compiler infrastructure

Most of the high-level power/performance optimization algorithms are implemented in the second stage of the PACT HDL compiler. First, it reads in the HDL AST and generates the CDFG graph. This graph is a hierarchical sequencing graph represented by basic blocks with its own local data flow graph and edges indicating the dependency relationship between them. After structural and data flow analysis, different optimization algorithms are applied aiming at different targets. IP core libraries can also be included for customized optimization. Finally, optimized HDL AST is generated.

Compiler Optimizations in the PACT HDL Behavioral Synthesis Tool

We have developed numerous compiler optimizations within the PACT compiler framework. We have developed techniques for building a Control and Data Flow Graph (CDFG) from a C program description using the SUIF framwork. We have developed numerous scheduling algorithms such as As Soon As Possible (ASAP), As Late As Possible (ALAP), Resource Constrained ALAP, and Resource Constrained ALAP algorithms. Furthermore we have developed compiler optimizations such as constant propagation, constant folding, common sub-expression elimination, pipelining, and dead code elimination on the CDFG representation. Several compiler algorithms for power and performance optimizations were presented. Experimental results show that for Constant PF and DE, an average of 15.9% power savings and 24.0% energy savings can be obtained for ASICs and 12.1% energy savings for FPGAs. The promising results show that the compiler optimizations can have a potential power savings and can gain good tradeoffs between power, energy, and performance. Figure 2 shows some results of these optimizations on ASICs. Figure 3 shows some results of these optimizations on FPGAs.

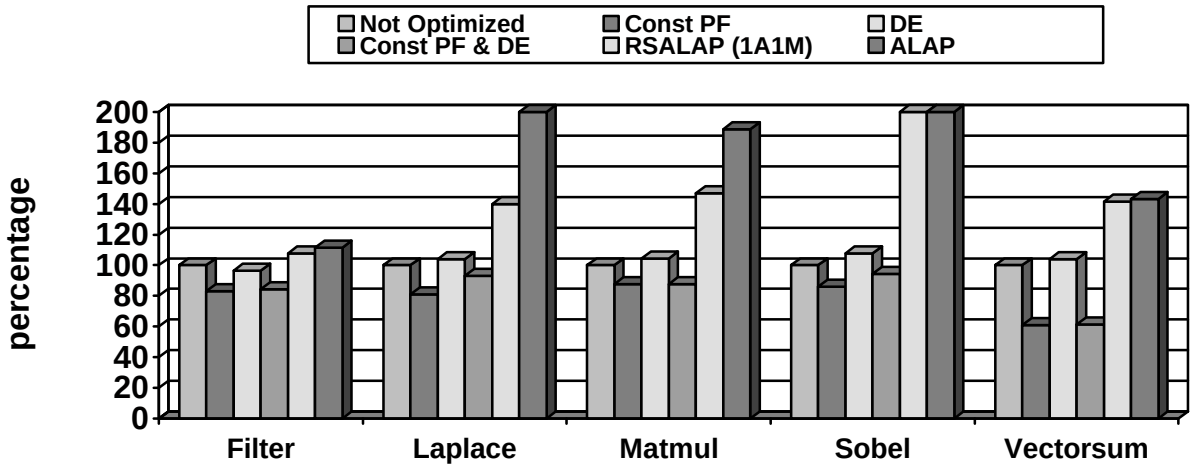

Figure. 2. ASIC Power Results (0.25 micron, 2.5 V)

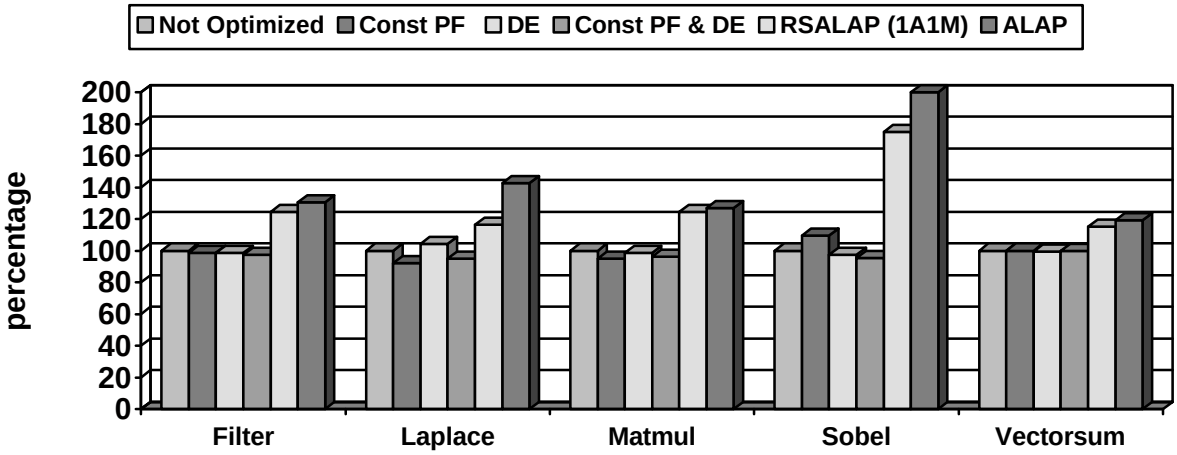

Figure. 3. FPGA Power Results (Xilinx XCV400 device, 2.5 V)

Development of Power Aware Synthesis and Scheduling Algorithms

We have developed the ETAIP algorithm for power aware synthesis of designs for FPGAs and ASICs. ETAIP (Energy constrained Timing optimization Algorithm with IP core library) is one of the power optimization algorithms developed in the PACT compiler to utilize the benefit of IP cores. The basic idea of the ETAIP algorithm is to start with an assignment of the minimum energy IP core for each operation in the DFG. Subsequently, the algorithm finds the critical paths in the DFG. For all nodes in the critical path, it computes the increase in energy (IE) consumed if the IP core is replaced

by a faster IP core for that operation. It also computes the decrease in cycle time (DC) if the IP core is replaced by a faster IP core for that operation. Among all the possible IP core replacements, the algorithm greedily selects the one which maximizes the IE/DC ratio. This step is repeated iteratively until there are no new IP cores found. The reason that we use the energy constraint instead of power constraint in the ETAIP algorithm is that we exploit clock-gating techniques and assume that the IP core does not consume power when it is inactive. Thus each IP core only consumes power during its execution cycles. A benchmark suite of four image processing codes and kernels was used to test and evaluate the ETAIP optimization algorithm withing the PACT compiler. The power and energy results of ETAIP algorithm for the innermost loop are displayed in Figure 4, Figure 5 respectively. The performance of the ETAIP9 algorithm is always better than that of ETAIP1, especially for the ones with more resources, as they can efficiently leverage the parallelism of the original C program. The power and energy of the ETAIP9 algorithm is always larger than those of ETAIP1 since ETAIP1 optimizes the output with the smallest energy IP core binding.

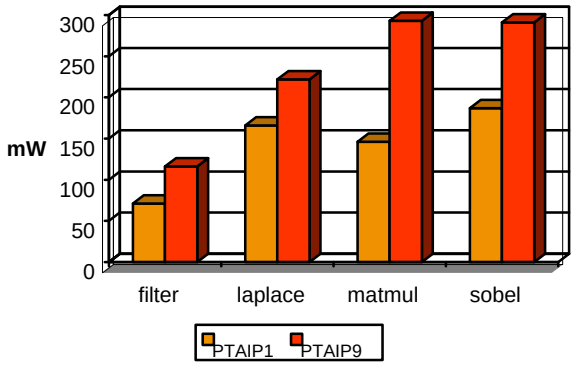

Figure 4. Power results for ETAIP

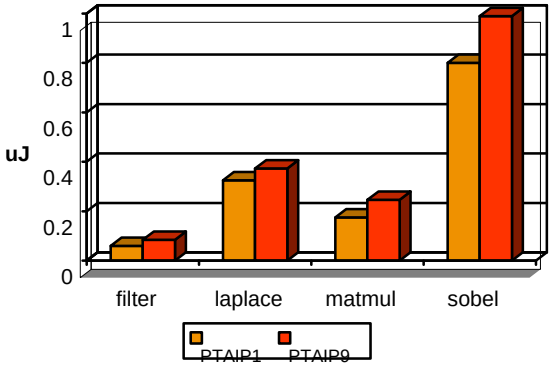

Figure 5: Energy results for ETAIP

Handling Data Streams While Compiling C Programs Onto Hardware

Recently, there have been several efforts to develop compilers that synthesize applications written in high-level languages such as C into hardware. Streaming

applications form an important class of systems that need to be mapped onto hardware. It is unclear how infinite data streams can be handled within the context of a C-based behavioral compiler. In this paper we present the modeling of stream and frame based applications in C and extend a behavioral synthesis tool framework to synthesize RTL models from the C descriptions. Individual C functions get synthesized onto individual IP blocks in the synthesis framework. This paper describes a system level approach for interfacing synthesized IP blocks. We describe how the behavioral synthesis tool automatically infers ports and generates the required handshaking signals for interfacing the IP blocks in the system and synthesizes top level of the system hierarchy. We also present ways to model parallelism among C functions which get translated to IPs working in parallel in the system. Results of the system level synthesis and simulation are presented for single IP systems for streaming data and frame-based data, and for multiple IP blocks having streaming and frame-based data.

Power Aware Interface Synthesis

A large portion of a system's power is consumed on the global on-chip busses. In this task, we addressed the problem of interface synthesis for a system-on-a-chip (SOC). We developed some techniques to minimize the power consumption on the interface while meeting the latency requirements. The focus was on bus-based designs using the AMBA protocol. An analysis of several bus-based designs was tried and the impact of the transformations on power and latency was investigated. Based on this analysis, a heuristic was proposed that addressed the problem of finding the number of layers for the bus and assigning components to each layer, so that the power is minimized. Results showed that the heuristic can produce a bus architecture that consumes less power than the single layer architecture.

1.7. Task C: Power-Aware CAD Tools

Research on power aware CAD tools has been pursued in two distinct directions

- Development of improved strategies for power optimization and management
- Development of improved design methodologies and design philosophies for better estimation and optimization

Development of improved strategies for power optimization and management

Power optimization and management are key to success of modern integrated systems which are becoming largely mobile and cost critical. We developed novel and innovative strategies/algorithms for addressing power. These include innovative strategies for voltage scaling and gate sizing simultaneously in gate level designs, optimization and management of power in clock trees which comprise of a very large chunk of overall power dissipation, quick and accurate estimation models for enabling early in optimization and estimation, strategies for power management in distributed systems like sensor networks. Each of these contributions are described in detail as follows

- Gate Level Power Optimization Using Supply Voltage Scaling: The work presents an approach for applying two supply voltages to optimize power in CMOS digital circuits under the timing constraints. Given a technology-mapped network, we first analyze the power/delay model and the timing slack distribution in the network. Then a new strategy is developed for timing-constrained optimization issues by making full use of slacks. Based on this strategy, the power reduction is translated into the polynomial-time-solvable maximal-weighted-independent set problem on transitive graphs. Since different supply voltages used in the circuit lead to totally different power consumption, we propose a fast heuristic approach to predict the optimum dual-supply voltages by looking at the lower bound of power consumption in the given circuit. To deal with the possible power penalty due to the level converters at the interface of different supply voltages, we use a "constrained F-M" algorithm to minimize the number of level converters. We have implemented our approach under SIS environment. Experiment shows that the resulting lower bound of power is tight for most circuits and that the predicted "optimum" supply voltages are exactly or very close to the best choice of actual ones. The total power saving of up to 26% (average of about 20%) is achieved without degrading the circuit performance, compared to the average power improvement of about 7% by gate sizing technique based on a standard cell library. Our technique provides the power delay tradeoff by specifying different timing constraints in circuits for power optimization. Another work considers simultaneous voltage scaling and gate sizing for optimizing power. The key idea here is to gain more slack by performing gate sizing and use that extra slack for optimizing power through voltage scaling. Algorithms and complexity issues are discussed.

- Activity Driven Clock Design: In this work, we investigate reducing the power consumption of a synchronous digital system by minimizing the total power consumed by the clock signals. A clock network that distributes the synchronizing signal can incorporate for a significant portion of the overall chip power (20%-40%). Hence effective methods of reducing clock power are required. This work proposes clock gating for reducing clock power. We construct activity-driven clock trees wherein sections of the clock tree are turned off by gating the clock signals. Since gating the clock signal implies that additional control signals and gates are needed, there exists a tradeoff between the amount of clock tree gating and the total power consumption of the clock tree. We exploit similarities in the switching activity of the clocked modules to reduce the number of clock gates. Assuming a given switching activity of the modules, we propose three novel activity-driven problems: a clock tree construction problem, a clock gate insertion problem, and a zero-skew clock gate insertion problem. The objective of these problems is to minimize system's power consumption by constructing an activity-driven clock tree. We propose an approximation algorithm based on recursive matching to solve the clock tree construction problem. We also propose an exact algorithm employing the dynamic programming paradigm to solve the gate insertion problems. Finally, we present experimental results that verify the

effectiveness of our approach. This paper is a step in understanding how high-level decisions (e.g., behavioral design) can affect a low-level design (e.g., clock design).

- Power Optimization in Distributed Systems: Up until now, low power system approaches have been restricted to single physical system. Recent emergence of distributed embedded systems has created a need for power optimization across individual system boundaries. In particular a great deal of excitement has been generated by wireless adhoc sensor networks which integrate communication, computation and sensing elements into self organizing, adaptive and multifunctional systems. We also addressed power management in such systems. Focusing our attention on node scheduling for minimum degree coverage, we develop provably optimal polynomial time algorithms. Furthermore we analyze the scaling properties of the problem as the number of sensors increase. Interesting insights and observations were obtained.

- Early On Power Estimation: Effectiveness of early on power optimization is well known. But this optimization is heavily dependent on the quality of early estimation. Hence it is highly imperative to develop accurate models that are fast to evaluate so that they can be effectively used in high level power optimization techniques. Our work is a step in this direction. This work concentrates on quick evaluation techniques to estimate the quality of resource binding from a power point of view without performing the computationally expensive binding step in High Level Synthesis. A design flow that takes C codes from Media Bench Suite and performs scheduling and power driven binding is constructed. An iterative rescheduling and rebinding step for power improvement follows this. The design flow integrates commercial tools like Synopsys, VSS and academic compilers like SUIF in a common optimization framework. The input to resource binder is a scheduled Date Flow Graph (DFG). This data flow graph has many properties like number of edges, their density etc. Based on these simple figures of merit, three metrics are proposed for evaluating the quality of a DFG from a power point of view. The metrics had correlation as high as 0.95 and more than 0.75 for most test cases when compared to post binding power values. Optimizing for these metrics results in schedules with power dissipation very close to schedules with minimum power after binding and rescheduling. Results also show that metric evaluation is on an average 42.6 times faster than optimal binding and iterative power improvement. Hence these metrics enables fast design exploration at early stages.

Development of improved design methodologies and design philosophies for better estimation and optimization

The previous aspect of our research addressed algorithms and strategies for power optimization/estimation and management. A very important aspect of power optimization in specifically and design automation in general lies with the design flow itself. Typical VLSI CAD flow comprises of High Level Synthesis followed by Logic Synthesis and concluded by Physical Design. During each of these design steps, a pertinent cost

function (like area, delay, power, wire-length etc) is optimized without any regard to its effect on other cost functions and overall design quality as a whole. In this research we embarked upon an ambitious project of re-evaluating the traditional design flow especially from a power point of view. We obtained some path-breaking results which are elaborated as follows

- **SLACK: The Universal Optimization Objective:** Slack is defined as the increase in delay on basic elements of a design (like gates or functional modules) which can be tolerated without violating the overall delay constraint. Higher the overall slack, better it is for design quality and design closure.

    - Slack a new optimization metric for gate level designs: Our work proposes potential slack as a new metric for evaluating the quality of gate level designs. Given a gate level design and a delay constraint, potential slack is defined as the sum total of the increase in delay that can be tolerated for the design without violating the delay constraint. This extra *tolerable* delay could then be used to optimize other design metrics like power, routability etc. visits the problem from an algorithmic point of view by modeling the gate level design as a *directed acyclic graph (DAG).* Notions of slack sensitivity and budget gradients are proposed to demonstrate the characteristics of the potential slack problem. We have evaluated the potential slack metric with various problems in VLSI CAD and demonstrates the effectiveness of potential slack in optimizing design metrics like power.

    - Optimization and management of slack during high level synthesis: In this aspect of our research we evaluated slack from a high level synthesis point of view. The first work in this regard addresses the scheduling problem. The scheduling problem associates a control step (clock step) to each operation of a data flow graph such that the timing and resource constraints are satisfied. In this work, we presented two scheduling approaches for DFGs. Firstly, an algorithm that produces schedules with maximum slack under no resource constraints is presented. Secondly the same problem is re-visited with resource availability as a constraint. Various theoretical aspects of the problem are evaluated.

      The approach evaluates the post scheduling slack optimization and management problem. Essentially algorithms like our previous ones will produce schedules with large operation slack. This extra slack could be used/managed for getting better design quality and design closure. We propose algorithms which transform the operation slack to relaxed delay constraints for functional modules at RT-Level. A new metric called delay relaxation parameter (DRP) for RTL designs is proposed. DRP essentially captures the degree of delay relaxation that the design can tolerate without violating the clock constraint. This metric when optimized results in quicker design flow. Algorithms to optimize DRP are formulated

and their optimality are investigated. These algorithms took a scheduled DFG as input and converted its operation slack into relaxed delay constraints. This was performed in two steps: delay budgeting and resource binding. First, the available slack after scheduling was converted into extra delay budgets for each operation without violating any of the data dependency or resource constraints. This extra operation delay was converted to relaxed delay constraints for functional modules through effective binding. Experimental results are conducted using a state of the art design flow with Synopsys Design Compiler followed by Cadence Place and Route. Our approach of optimizing DRP resulted in lesser design iterations and faster design closure as compared to designs generated through Synopsys Behavioral Compiler and a representative academic design flow. For most benchmarks, our approach resulted in faster clock frequency when compared with other approaches. Even when both approaches met the clock constraint, our approach was better in design quality (placement area, routing wirelength, number of vias). The overall design time was faster. These comparisons were made at layout level, therefore they represent real improvements. The relaxed delay constraints could also be used to optimize power through techniques like voltage scaling.

- **Predictability: Definition Analysis and Optimization:** The most significant contribution of this research has been the development of the philosophy of predictability applied to power optimization and estimation. System level power prediction is extremely hard due to the different and complex parameters involved. This research seeks to approach predictability by quantification of the error associated with estimation. We believe, if the cost function must be optimized, so should be the error. Such would be a predictability driven approach. Through our experiments we found that power predictability is strongly associated with the design structure and specifically resource binding. We propose, variance, as an effective measure for capturing predictability, hence an important optimization objective. Two approaches for optimizing power predictability are proposed. This philosophy of predictability was further advanced by generating tradeoff curves between power and predictability in RTL designs. Optimal pseudo-polynomial algorithms and approximation algorithms based on the knapsack problem were proposed and investigated.

    - Predictability driven binding: The work addresses the problem of optimizing the predictability of power estimation in RTL designs through effective resource binding. Predictability is defined as the quantified for of accuracy. A predictability driven approach tries to optimize/maximize the predictability of various cost functions through different design decisions. Such an approach also tries to model cost functions as probability distributions in the presence of uncertainties and optimizes the likelihood of satisfying a delay constraint. A predictability driven approach will have many key subcomponents: estimation of future uncertainties, optimization

of predictability, optimization of design constraint satisfiability are important ones among them. In this work we address the problem of predictability of power estimation in RTL designs. We propose effective resource binding techniques to solve the problem. The resource binding problem is modeled as a Min-Cost flow formulation on a compatibility graph. The nodes of this graph represent operations. Any two operations that can be potentially bound on the same resource have an edge between them. The cost on these edges signifies the unpredictability in power estimation if the corresponding operations are bound together. The Mincost formulation tries to minimize the overall unpredictability in the binding solution. This work was further extended to minimize the variance of the power distribution in RTL designs. To test our ideas, an experimental framework is built by integrated academic compilers like SUIF along with commercial tools like Synopsys Design Compiler and the VSS Simulator. Comparisons were made with traditional low power driven resource binding approach. Results showed that our approach was very effective in optimizing predictability (84% on average) at minimal power penalty. This increased predictability could be used to improve the quality of system level power management techniques.

- Predictability in RTL Designs: This work addressed the predictability of power dissipation in RTL designs but the approach is very different from earlier work. In this work we observed that if the delay constraint is stringent then the predictability in power dissipation is higher. This seems to be the case because a stringent delay constraint provides very little flexibility to low level tools hence the design space is severely constrained. This increases the predictability in power estimation although it comes with an increase in power dissipation. We present optimal pseudo-polynomial algorithm to optimize predictability through this strategy while keeping the power dissipation within a pre-specified budget. We further extend this approach and propose an approximation algorithm. Experimental results showed the effectiveness of this approach.

The big picture includes development of a complete predictability driven design methodology that not only optimizes the design quality but also the associated error. Such an approach would be more qualified to address the modern deep sub-micron problems especially from a power perspective.

1.8. List of professional personnel supported on this project

- PROF. PRITHVIRAJ BANERJEE, Walter Murphy Professor, Electrical and Computer Engineering and Center for Parallel and Distributed Computing, Northwestern University

- PROF. ALOK CHOUDHARY, Professor, Electrical and Computer Engineering and Center for Parallel and Distributed Computing, Northwestern University.

- PROF. ANDREAS MOSHOVOS, Assistant Professor, Electrical and Computer Engineering and Center for Parallel and Distributed Computing, Northwestern University .

- PROF. MAJID SARRAFZADEH, Professor, Computer Science, UCLA

- Kohinoor Basu, M.S. student, ECE, NWU, Worked on PACT architecture (advisor A. Choudhary and A. Moshovos)

- Gaurav Mittal, Ph.D. student, ECE, NWU, Worked on PACT architecture (advisor A. Moshovos)

- Amirali Beniasadi, Ph.D. student, ECE, NWU, Worked on PACT architecture (advisor A. Moshovos)

- Jay Prakash Pisharath, Ph.D. student, ECE, NWU, Worked on PACT architecture (Advisor Alok Choudhary)

- Joe Zambreno, Ph.D. student, ECE, NWU, Worked on PACT Architecture (Advisor Alok Choudhary)

- Satrajit Pal, M.S. student, ECE, NWU, Worked on PACT compiler (advisor P. Banerjee and A. Choudhary)

- Debabrata Bagchi, M.S. student, ECE, NWU, Worked on PACT compiler (advisor A. Choudhary and P. Banerjee)

- Alex Jones, Ph.D. student, ECE, NWU, Worked on PACT compiler (advisor P. Banerjee)

- Tianyi Jiang, Ph.D. student, ECE, NWU, Worked on PACT compiler (advisor P. Banerjee)

- Xiaoyong Tang, Ph.D. student, ECE, NWU, Worked on PACT compiler (advisor P. Banerjee)

- Nikos Liveris, Ph.D. student, ECE, NWU, Worked on PACT compiler (advisor P. Banerjee)

- Ankur Srivastava, Ph.D. student, CS, UCLA, Worked on PACT CAD Tools (advisor M. Sarrafzadeh)

- Ryan Kastner, Ph.D. student, CS, UCLA, Worked on PACT CAD Tools (advisor M. Sarrafzadeh)

1.9. List of written publications in journals or reports (June 2000 to May 2003)

- Chunhong Chen, Ankur Srivastava and Majid Sarrafzadeh, “On Gate-Level Power Optimization Using Dual Supply Voltages”, IEEE Transactions on Very Large Scale Integrated Systems, vol. 9, pp. 616-629, Oct 2001. (TVLSI)

- A.H. Farrahi, C. Chen, A. Srivastava, M. Sarrafzadeh and G. Tellez, "Activity Driven Clock Design ", IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, vol. 20, pp. 705-714, Jun 2001. (TCAD)

- Chunhong Chen and Majid Sarrafzadeh, "Simultaneous Voltage Scaling and Gate Sizing for Low Power Design," submitted to IEEE Transactions on CAD.

- Chunhong Chen, Xioajian Yang, and Majid Sarrafzadeh, "Predicting Potential Performance for Digital Circuits," IEEE Transactions on CAD, Vol. 21, No. 3, March 2002, pp. 253-262

- Chunhong Chen, Elaheh Bozorgzadeh, Ankur Srivatsava, and Majid Sarrafzadeh, "Budget Management and Its Applications," Algorithmica, Vol 34, No-3, pp. 261-275.

- P. Banerjee, V. Saxena, J. Uribe, M. Haldar, A. Nayak, V. Kim, S. Parkes, D. Bagchi, S. Pal, D. Zaretsky, N. Tripathi, B. Jiang, R. Anderson, T. Vanevenhoven, D. Nandy, "Overview of a Compiler for Synthesizing MATLAB Programs onto FPGAs," Submitted to IEEE Transactions on VLSI Systems.

1.10. Papers presented at meetings, conferences, seminars (June 2000 to May 2003)

- Nan Jiang, Jayaprakash Pisharath, and Alok Choudhary. Characterizing and Improving Energy-Delay Tradeoffs in Heterogeneous Communication Systems. To appear in Proc. of the International Symposium on Signals, Circuits and Systems (SCS), July 2003, IEEE Press.

- Jayaprakash Pisharath, Nan Jiang, and Alok Choudhary. Evaluation of Application-Aware Heterogeneous Embedded Systems for Performance and Energy Consumption. In Proc. of the 9th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), May 2003, IEEE Press.

- Kohinoor Basu, Alok Choudhary, Jayaprakash Pisharath, and Mahmut Kandemir. Power Protocol: Reducing Power Dissipation on Off-Chip Data Buses. In Proc. of the 35th Annual International Symposium on Microarchitecture (Micro), November 2002, Istanbul, Turkey, IEEE Press.

- Jayaprakash Pisharath and Alok Choudhary. An Integrated Approach to Reducing Power Dissipation in Memory Hierarchies. In Proc. of the International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES), October 2002, ACM Press.

- A. Nayak, M. Haldar, P. Banerjee, C. Chen, and M. Sarrafzadeh, "Power Optimization of Delay Constrained Circuits," Proc. Application Specific Integrated Circuit/System-on-a-Chip Conference (ASCI/SOC 2000), Washington, DC, September 2000.

- V. Kim, P. Banerjee, K. De, and J. Brouwers, "Parallel and Distributed VLSI Synthesis for Commercial CAD on a Network of Workstations," Proc. 12th IASTED International Conference on Parallel and Distributed Computing Systems (PDCS 2000), Las Vegas, NV, November 6-9, 2000.

- M. Haldar, A. Nayak, A. Choudhary, and P. Banerjee, "Scheduling Alogorithms for Automated Synthesis of Pipelined Designs on FPGAs for Applications Described in MATLAB", Proc. International Conference on Compilers, Architectures and Synthesis for Embedded Systems (CASES 2000), Nov. 2000, San Jose, CA.

- M. Haldar, A. Nayak, N. Shenoy, A. Choudhary, and P. Banerjee, "FPGA Hardware Synthesis from MATLAB," Proc. of VLSI Design Conf. Jan. 2001, Bangalore, India.

- N. Shenoy, P. Banerjee, A. Choudhary, and M. Kandemir, "Efficient Synthesis of Array Intensive Computations onto FPGA Based Accelerators," Proc. of VLSI Design Conf. Jan. 2001, Bangalore, India.

- M. Haldar, A. Nayak, A. Choudhary, and P. Banerjee, "Automated Synthesis of Pipelined Designs on FPGAs for Signal and Image Processing Applications Described in MATLAB," Proc. Asia Pacific Design Automation Conf (ASP-DAC), Feb. 2001, Tokyo, Japan.

- M. Haldar, A. Nayak, A. Choudhary, and P. Banerjee, ""FPGA Hardware Synthesis from MATLAB Utilizing Optimized IP Cores" Proc. Ninth ACM/SIGDA International Symposium on Field Programmable Gate Arrays., Feb. 2001, San Jose, CA.

- A. Nayak, M. Haldar, A. Choudhary, P. Banerjee, "Precision And Error Analysis Of MATLAB Applications During Automated Hardware Synthesis for FPGAs," Proc. Design Automation and Test in Europe (DATE 2001), Mar. 2001, Berlin, Germany.

- A. Nayak, M. Haldar, A. Choudhary and P. Banerjee, "Parallelization of MATLAB Applications for a Multi-FPGA System," Proc. FPGA Symp. on Custom Computing Machines (FCCM-2001), Napa Valley, CA, Apr. 2001.

- A. K. Jones, and P. Banerjee, "Parallel Implementation of Matrix and Signal Processing Libraries on FPGAs," Proc. IASTED Parallel and Distributed Computing Systems Conf. (PDCS2001), Anaheim, CA, Aug. 2001.

- P. Banerjee, M. Haldar, A. Nayak, A. Choudhary, "Overview of the MATCH Compiler for Compiling MATLAB Programs into Hardware," Proc. of NASA Earth Science Technology Conference, Aug. 2001, Washington, DC.

- M. Haldar, A. Nayak, A. Choudhary, P. Banerjee, "A System for Synthesizing Optimized FPGA Hardware from MATLAB," Proc. Int. Conf. on Computer Aided Design, Nov. 4-8, 2001, San Jose, CA.

- A. Nayak, M. Haldar, A. Choudhary, and P. Banerjee, "Accurate Area and Delay Estimators for FPGAs," Proc. Design Automation and Test in Europe (DATE-2002), Mar. 2002, Paris, France.

- A. Jones, D. Bagchi, S. Pal, X. Tang, A. Choudhary, and P. Banerjee, "PACT HDL: A C Compiler with Power and Performance Optimizations," Proc. International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES 2002), Grenoble, France, October 2002.

- P. Banerjee, M. Haldar, A. Nayak, V. Kim, D. Bagchi, S. Pal, N. Tripathi, "A Behavioral Synthesis Tool For Exploiting Fine Grain Parallelism in FPGAs," Proc. International Workshop on Distributed Computing (IWDC), Dec. 28-30, 2002, Kolkata, INDIA. To Appear as Springer Verlag Lectures in Computer Science Series.

- P. Banerjee, "An Overview of a Compiler for Mapping MATLAB Programs onto FPGAs," Invited Paper at the Asia Pacific Design Automation Conference (ASP-DAC03), Jan. 2003, Japan.

- A. K. Jones, P. Banerjee. An Automated and Power-Aware Framework for Utilization of IP Cores in Hardware Generated from C Descriptions Targeting FPGAs, Proc. FPGA based Custom Computing Machines (FCCM) (poster paper), Apr. 2003, Monterey, CA.

- X. Tang, T. Jiang, A. K. Jones, P. Banerjee, "Compiler Optimizations in the PACT HDL Behavioral Synthesis Tool for ASICs and FPGAs," Proc. IEEE System on a Chip Conference, Portland, OR., Sep. 2003.

- T. Jiang, X. Tang, A. K. Jones, P. Banerjee, “Optimizing Power While Exploiting Fine Grain Parallelism in FPGAs” Proc. Int. Conf. Parallel and Distributed Computing Systems (PDCS03), Marina Del Rey, CA, Nov. 2003.

- R. Mukherjee, A. K. Jones, P. Banerjee, “System Level Synthesis of Multiple IP Blocks in the PACT Compiler,” Proc. Int. Conf. Parallel and Distributed Computing Systems (PDCS03), Marina Del Rey, CA, Nov. 2003

- S. Roy, D. Sinha and P. Banerjee, “An Algorithm for Trading Off Quantization Error with Hardware Resources for MATLAB based Hardware Design,” Submitted to the IEEE FPGA Conference, Feb. 2004.

- R. Mukherjee, A. K. Jones, P. Banerjee, “Handling Data Streams While Compiling C Programs Onto Hardware” Submitted to the Design Automation and Test in Europe (DATE 2004), Paris, France, Feb. 2004.

- G. Mittal, D. Zaretsky, R. Churchwell, X. Tang, and P. Banerjee, "Automatic Translation of Software Binaries onto FPGAs," Submitted to the Design Automation and Test in Europe (DATE 2004), Paris, France, Feb. 2004.

- N. Liveris, P. Banerjee, “Power Aware Interface Synthesis for Bus Based SOC Design,” Submitted to the Design Automation and Test in Europe (DATE 2004), Paris, France, Feb. 2004.
-
- Ankur Srivastava, J. Sobaje, Miodrag Potkonjak and Majid Sarrafzadeh, "Optimal Node Scheduling for Effective Energy Usage in Sensor Networks", IEEE Workshop on Integrated Management of Power Aware Communications, Computing and Networking 2002.

- Ankur Srivastava, J. Sobaje, Miodrag Potkonjak and Majid Sarrafzadeh, "Optimal Node Scheduling for Effective Energy Usage in Sensor Networks", System Level Power Optimization for Wireless Multimedia Communication, R. Karri and D. Goodman Edited, Kulwer Academic Publishers

- Eren Kursun, Ankur Srivastava, Seda Ogrenci Memik and Majid Sarrafzadeh, “Early Evaluation Techniques for Low Power Binding”, ISLPED August-2002, pp. 160-165

- Seda Ogrenci-Memik, Ankur Srivastava and Majid Sarrafzadeh, "Algorithmic Aspects of Uncertainty Driven Scheduling", IEEE International Symposium on Circuits and Systems 2002. (ISCAS)

- Ankur Srivastava, Seda Ogrenci Memik, Bo-Kyung Choi and Majid Sarrafzadeh, “Achieving Design Closure Through Delay relaxation Parameter”, To Appear in International Conference on Computer Aided Design 2003

- Ankur Srivastava, Eren Kursun, and M. Sarrafzadeh, “Predictability in RT-Level Designs”, Journal of Circuits, Systems and Computers, " special issue on Low Power IC Designs., Vol 11 No-4, August 2002, pp. 323-333.

- Ankur Srivastava, Majid Sarrafzadeh, “Predictability: Definition, Analysis and Optimization”, ICCAD, Nov 2002, pp 118-121.